Hayal Dargin

Final Report:

Profit Analysis

Problem Statement

Our customer, the superstore, thinks that there are uncertainties and mismanagement in their business decision making. There is fluctuation in their profit even if their sales are keep increasing. They want to increase their profits and they are asking us to find out what they should pay attention to in order to increase. The data set they have provided for me to work with has 20 variables listed below:

- Order ID
- Order Date
- Ship Date
- Ship Mode
- Customer ID
- Customer Name
- Segment
- Country
- City
- State
- Postal Code
- Region
- Product ID
- Category
- Sub-Category
- Product Name
- Sales
- Quantity
- Discount
- Profit

In order to find out reasons behind decrease in profit First, I need to do a deep data analysis to increase the profit. What variables are affected by the profit most and in what way it will lead me to possible erroneous management and advertising decisions. There are possible questions to ask my data to gain more insights such as :

- Analysis of various business entities can be:
    - Customer Analysis (What my customer is buying from us the most?)
    - Sales Analysis (What is the highest sales in terms of time, product, area)
    - Product Analysis (Which product is selling more)
- Analysis of the above entities across dimensions can be:
    - Geographical Hierarchy
    - Product Hierarchy
- Analysis of Key Performance Indicators can be:
    - Sales
    - Profits

By using Super Store data set, I created a tool to help my customer in predicting their profit most impacted features to make better business decisions in their management and advertisements.

Data Wrangling

The row data from Super Store contained 9994 rows and 21 columns. While the data set of sales was usable, it had columns that would not be used in my analysis. I started looking at the data types and null values if any. There are no null values in my data set which is good. There were some columns that I did not need to keep in my data set therefore I dropped them such as Row ID. The data set is all about United States by region.

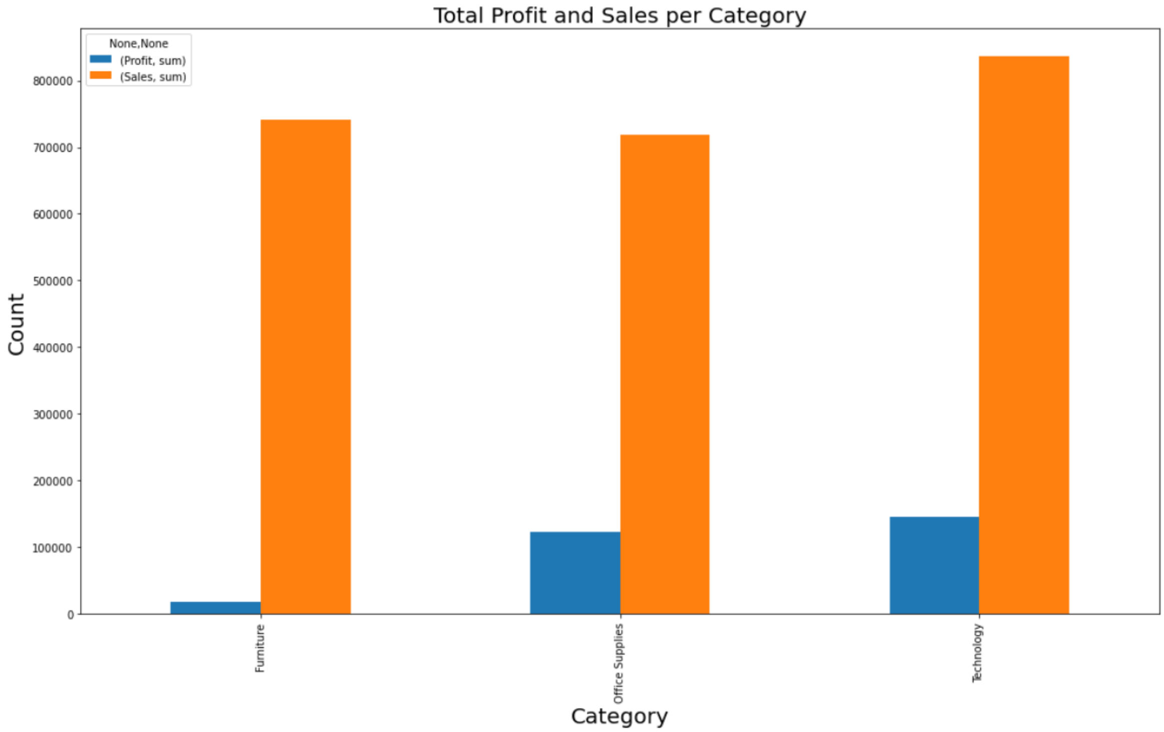
Since my data set was pretty clean, it did not require massive cleaning.

The final shape of my data set of Super Store was 9994 rows with 15 columns,

Exploratory Data Analysis

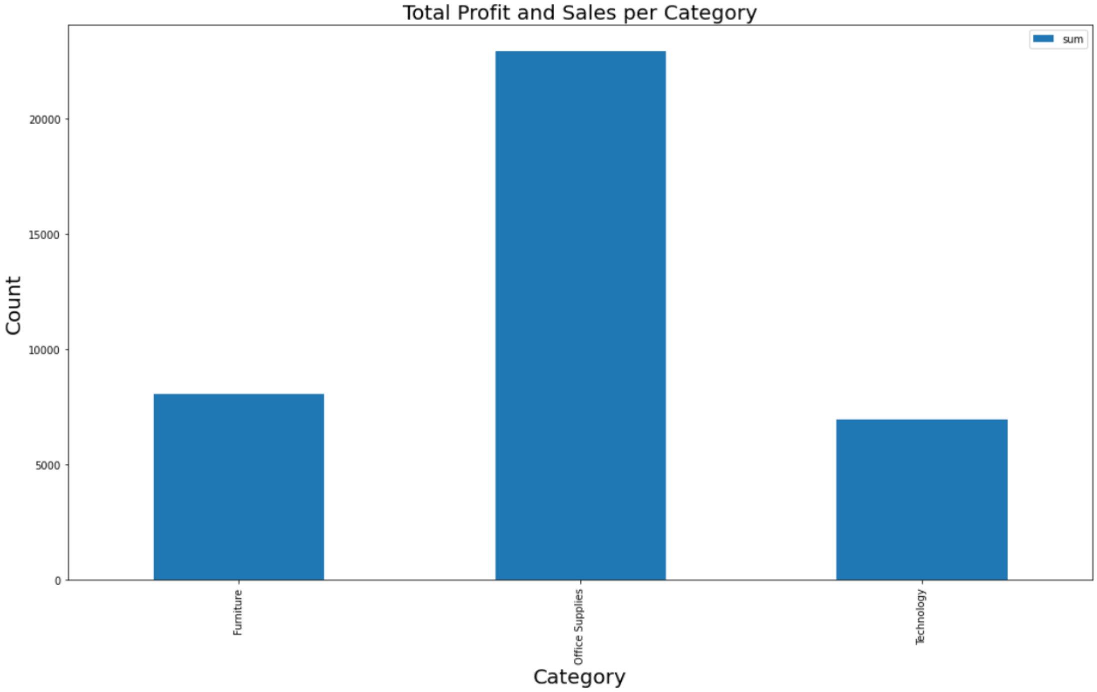
In data analysis, first I wanted to do category-based analysis by checking the different product categories. Looking at the profit and sales per category gave me technology followed by office supplies (figure 1).

Figure 1



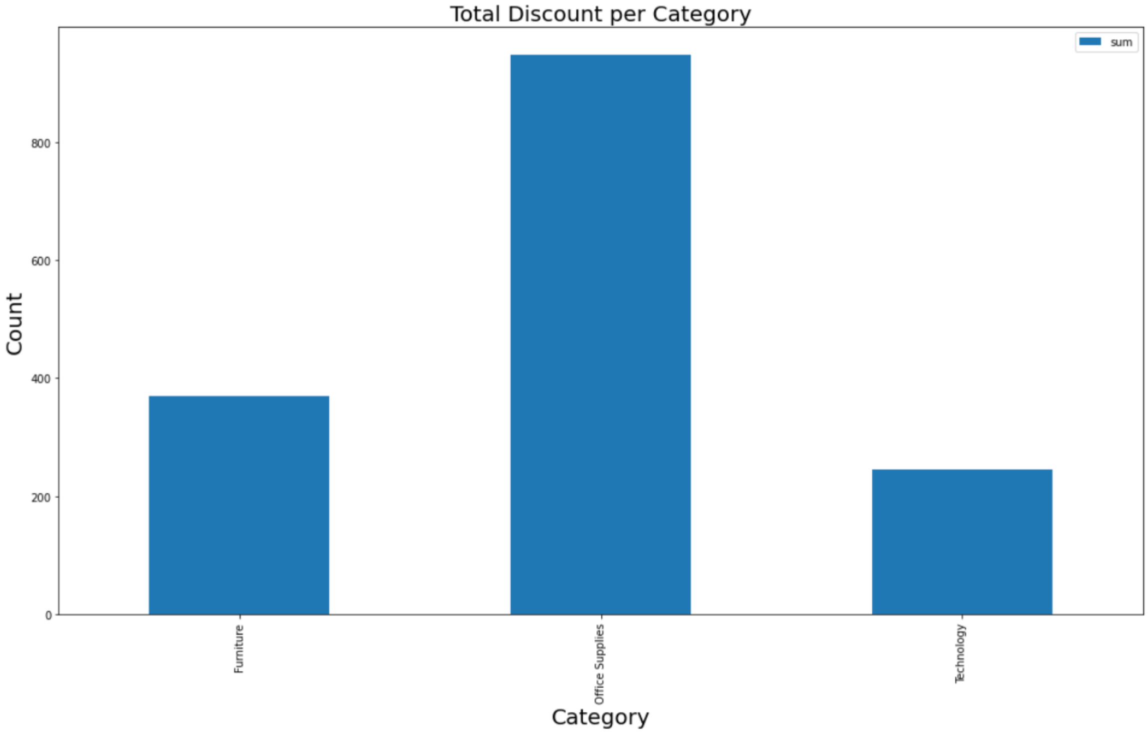
Looking at the total profit and sales per category showed high numbers in office supplies followed by furniture (figure 2). This tells us that price is having high impact over the profit.

Figure 2



When we go into total given discount per category, I found that the most discount is going to office supplies (figure 3).

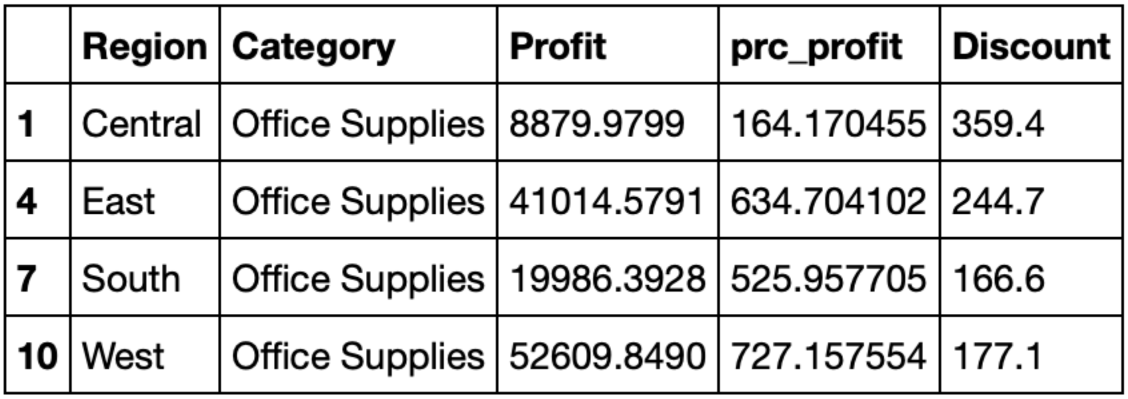
Figure 3



At the end of my general category analysis, I realized that although the highest profit comes from technology products, the most discounted products are office supplies products.

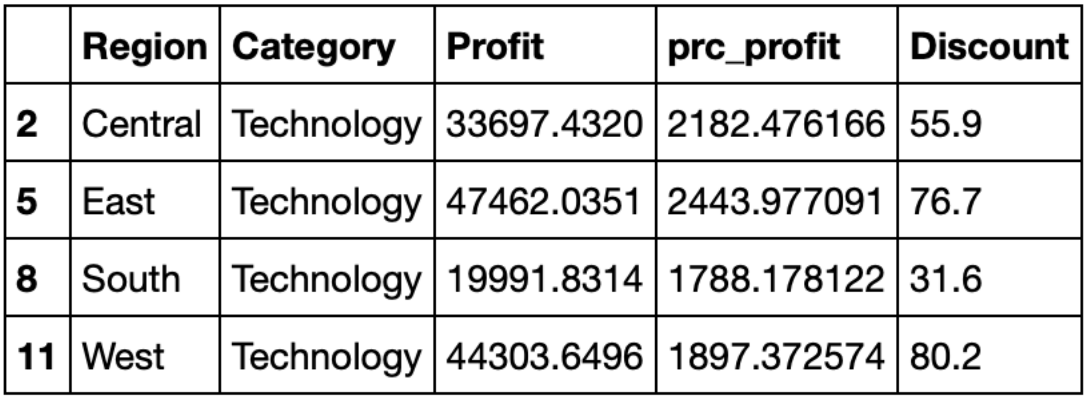
I then looked at the total count of sales per region and found that the west region is buying more than all other three regions. Also, east region is the second most buying region. This was a unique finding since the highest discount was given to the central region in office supplies and west region in technology (figure 4 and 5).

Figure 4



	Region	Category	Profit	prc_profit	Discount
1	Central	Office Supplies	8879.9799	164.170455	359.4
4	East	Office Supplies	41014.5791	634.704102	244.7
7	South	Office Supplies	19986.3928	525.957705	166.6
10	West	Office Supplies	52609.8490	727.157554	177.1

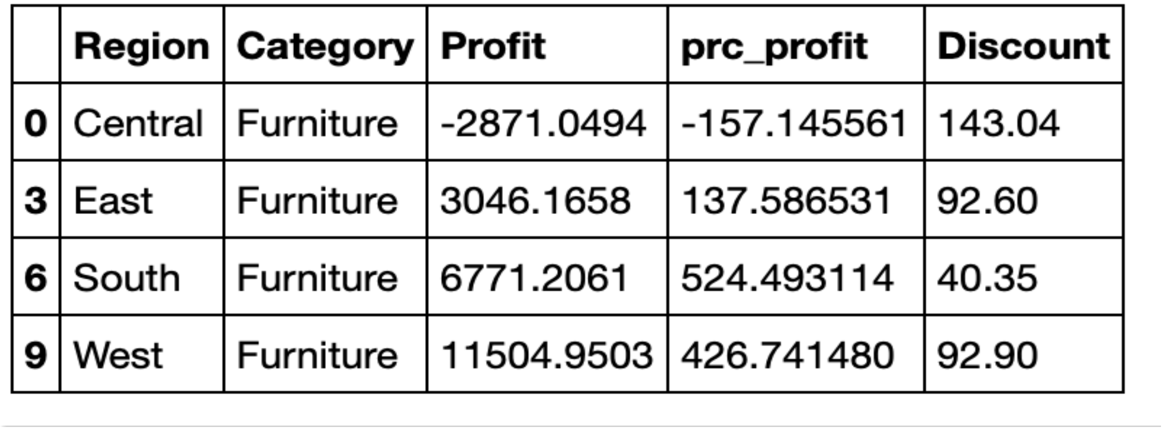
Figure 5



	Region	Category	Profit	prc_profit	Discount
2	Central	Technology	33697.4320	2182.476166	55.9
5	East	Technology	47462.0351	2443.977091	76.7
8	South	Technology	19991.8314	1788.178122	31.6
11	West	Technology	44303.6496	1897.372574	80.2

When I look at the furniture category, I see that the central region, where the biggest discount is given, has done a great deal of damage to the business. But following this, I saw that it brought the highest profit to the business despite the lowest discount given to the south region. As a result of this analysis, I can see very clearly that a big advertising error is being made in terms of discount per category and region(figure 6).

Figure 6



	Region	Category	Profit	prc_profit	Discount
0	Central	Furniture	-2871.0494	-157.145561	143.04
3	East	Furniture	3046.1658	137.586531	92.60
6	South	Furniture	6771.2061	524.493114	40.35
9	West	Furniture	11504.9503	426.741480	92.90

As a result of all analysis, I see that the increase/decrease in profit rate is not mainly related to sales but the discount. On the basis of years, although the sales amounts and numbers increased, the rate of profit did not increase in the same way. Even though the sales rates continued to increase in 2017, the profit rate decreased. Despite the increase in sales, the reasons for the decrease in the rate of profit may be the reduction given, advertising to the wrong regions, giving weight to the products in the wrong category, or increasing the cost (figure 7 and 8).

Figure 7

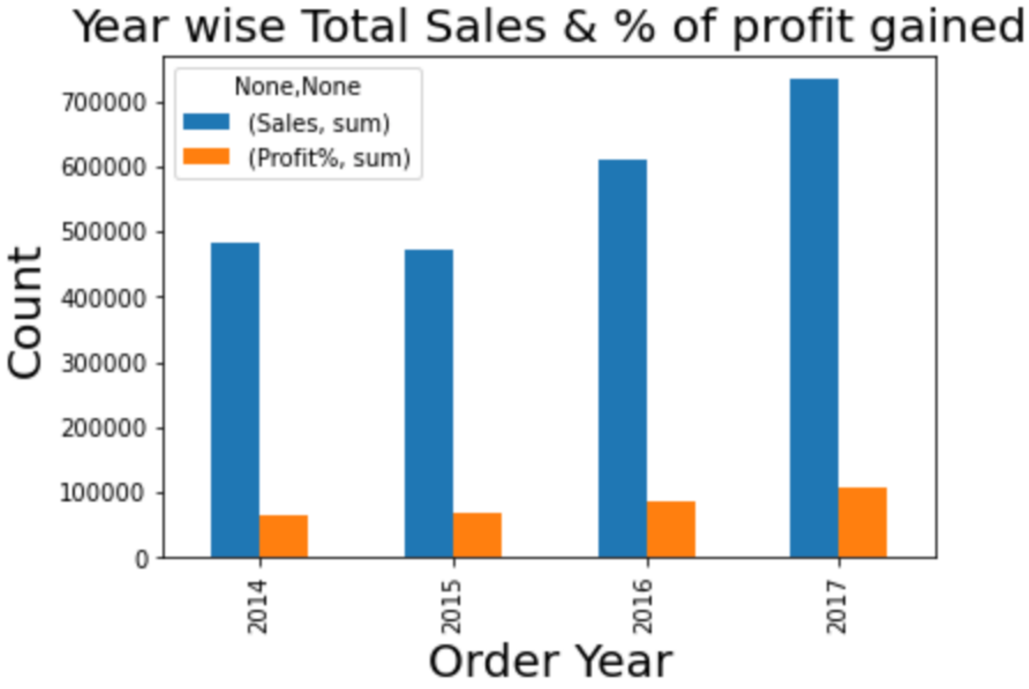
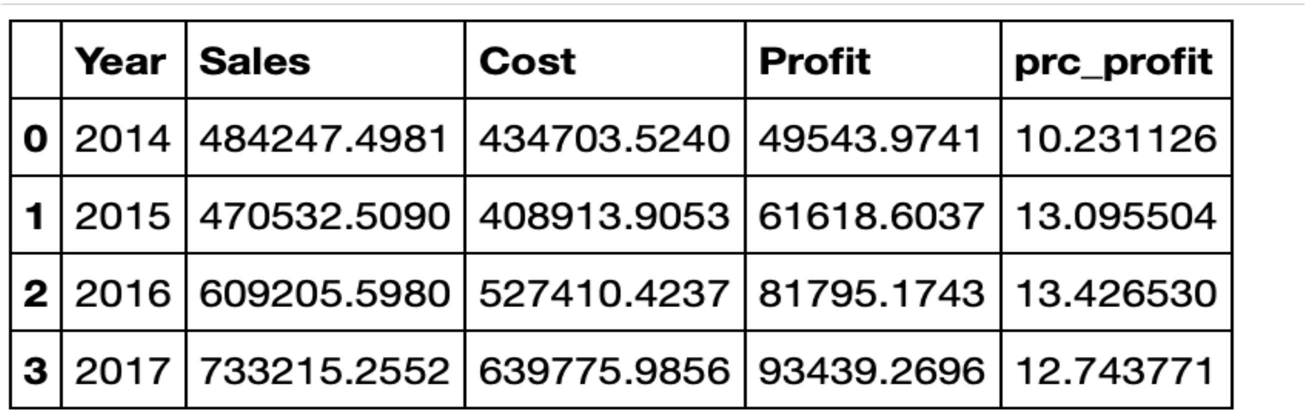


Figure 8



	Year	Sales	Cost	Profit	prc_profit
0	2014	484247.4981	434703.5240	49543.9741	10.231126
1	2015	470532.5090	408913.9053	61618.6037	13.095504
2	2016	609205.5980	527410.4237	81795.1743	13.426530
3	2017	733215.2552	639775.9856	93439.2696	12.743771

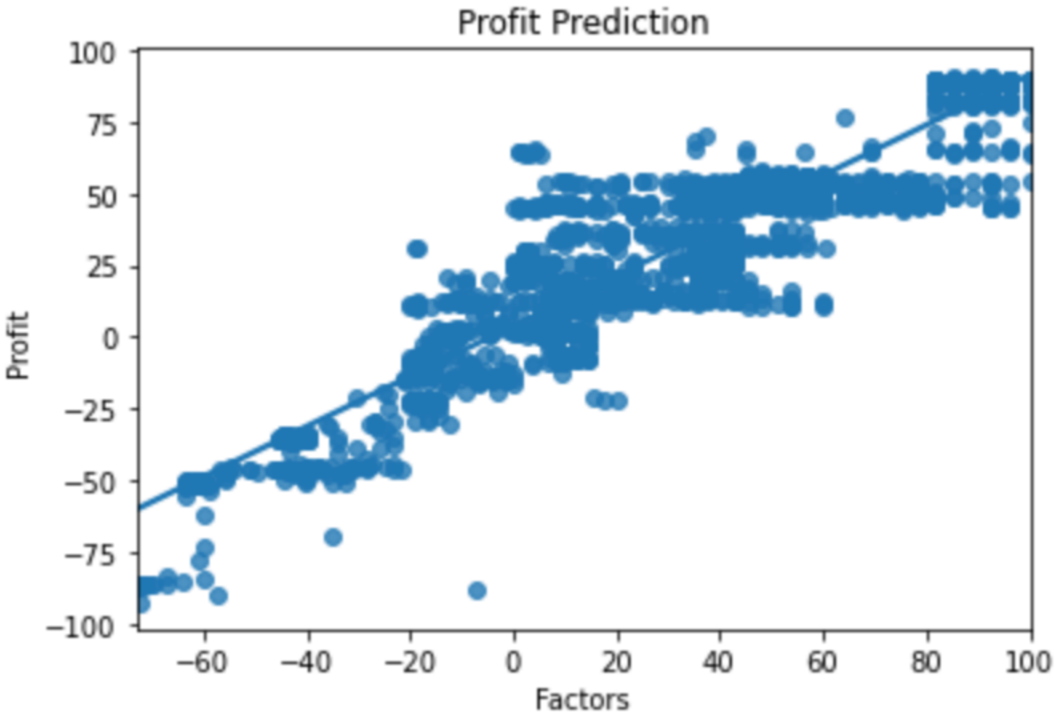
Model Selection

For modeling, I tested 2 different machine leaning regression models: Random Forest Regressor and Linear Regression. The metric I focused on when building my models were RMSE (root mean squared error), R Squared Score, MAE (mean absolute error) and model accuracy. I wanted my models to predict features with high correlation and importance.

Before building my models, however, I needed to do some feature selection. To do this I dropped some unwanted columns. Then first I encoded my categorical values to prepare my data set for modeling. Second, I used one-hot encoding on values which gives label in binary format, it is in terms of n digits where n is the number of classes, all digits are 0 except the class it represents which is 1.

I called linear regression model and passed in my training data to train the linear regression model. Evaluated my model's performance by using the RMSE (root mean squared error) and model accuracy. Profit Prediction is a graph of test values and predicted values. The dots represent the test values, and the line is of predicted values. We can see that there is a linear trend in the test values and our regression line captures it quite well. Hence, our model gives high accuracy of more than 87% but we can see there are outliers which reduce accuracy (figure 9).

Figure 9

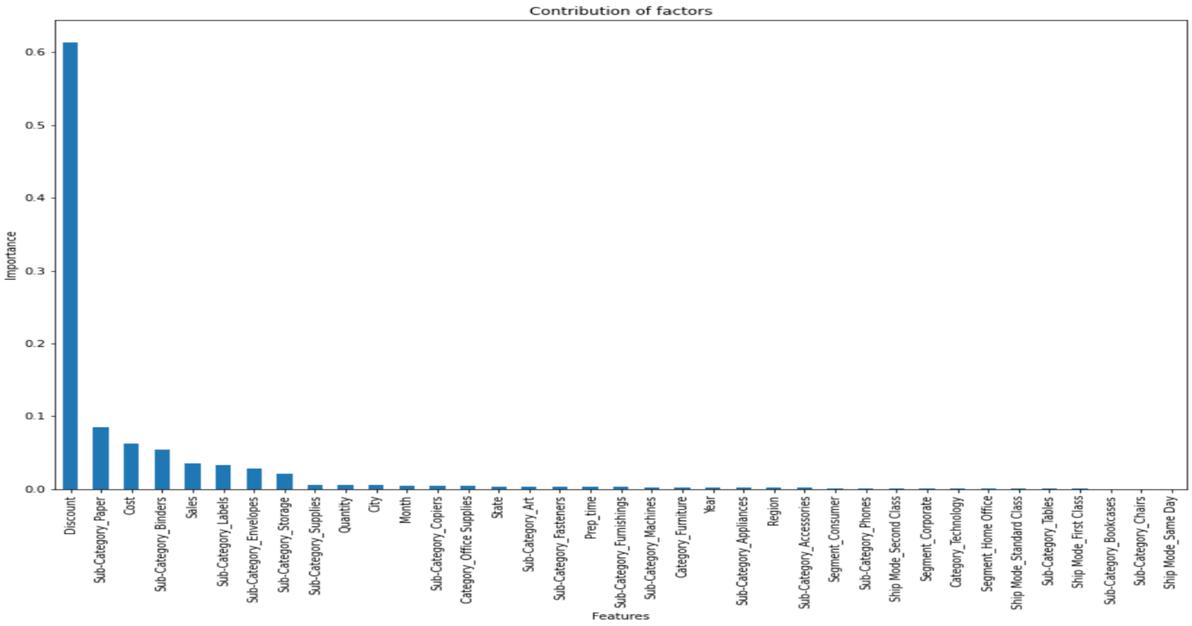


In order to select the best model, I evaluated my random forest model by rerunning the model with different numbers of estimators. I finalized my evaluation of random forest model by using the model accuracy metric and Random Forest is giving better accuracy level of 96% therefore I used random forest regressor as my final model with higher accuracy.

Conclusion

Looking at the last plot (figure 10), I can see that discount has the highest impact on profit with %61 importance level. This finding is expected, and I found that discount had a negative impact on sales and profit from the analysis. I had discount and profit analysis in the EDA section. In my analysis on category basis, although the highest discount is office supplies is $947.00, the profit is $122490.8008, the highest profit is achieved in technology products by $145454.9481 with lower discount of $244.40. When I looked at the furniture category on a region basis, I saw that business was in loss in the Central region by $-2871.0494, where the highest discount was given by 359.4 gave us the lowest profit of $8879.979. As a result, it is obvious that the discounts given are the most important variable but in negative direction.

Figure 10



Future Research

This can be taken into further research and analysis by looking at given discounts. While giving discounts, product category, region and demand should be taken into consideration.